Limits of Standard Deviation under constraints

Ian Hussey

Overview

This document provides a mathematical description of a constrained procedural solution for finding the minimum and maximum possible Standard Deviations for a truncated, granular variable (e.g., Likert scales), as implemented in the `tides_modified_from_.sd_limits()` R function.

This function was modified from Lukas Wallrich's `.sd_limits()` function included in the {rsprite2} package (version 0.2.1) after the original was found to produce false-negative results under some conditions. See the `demonstrate issues in .sd_limits.Rmd` file for a demonstration of these issues. Users should be aware of this potential for false-negative results with {rsprite2}.

The method employs the following constraints to calculate these limits on Standard Deviation:

- A specific mean ($\bar{x}$) and sample size n. Note that ($\bar{x}$) has typically been already rounded to some known degree of precision for reporting. This method accommodates this use of a rounded mean.
- A discrete measurement scale bounded by ℓ and u, as in Likert scales.
- The granularity of the scale. That is, its possible fractional increments, assuming the original data was either integer responses (e.g., a single-item Likert scale, or the sum score of a multi-item Likert scale) or the average of integer responses (e.g., a mean-scored Likert scale with k number of items).

Note that a closed form mathematical solution for this problem is also possible (see `Hussey & Cummins - 2025 - Limits of Standard Deviation under constraints - closed form solution.pdf`), although our current implementations of this solution produce false-positive results for as yet undetermined reasons. I therefore employ the method described here, and implemented in `tides_modified_from_.sd_limits()` in the {tides} package and recommend others to use this implementation.

Methodological Steps

1. Defining the Total Sum

The target total sum, total, is calculated as follows:

$$\text{total} = \frac{\text{round}(\bar{x} \times n \times k)}{k}. \tag{1}$$

This ensures that total is consistent with the discrete increments allowed by the scale.

2. Identifying Candidate Values α and β

The mean being tested is typically a reported value rounded to a certain level of precision. Choosing values exactly at ℓ and u might not allow the total sum to match the mean when accounting for rounding. Instead, two values, α and β are calculated. To find the distributions that yield extreme (minimum or maximum) Standard Deviations, the observations are approximately equally distributed over at most two distinct values α and β, which are chosen from:

$$\alpha, \beta \in \{\ell, \ell + \tfrac{1}{k}, \ldots, u\}. \tag{2}$$

- For maximum Standard Deviation, α is chosen near ℓ and β near u, the points at maximum possible distance from one another. The number of observations at the lower point, k, is placed at β, and the remaining $n - k$ observations are placed at upper point α.
- For minimum Standard Deviation, α and β are chosen as close together as the scale's granularity allows. The allocation of points follows the same logic, ensuring the total sum aligns with $\bar{x}$.

3. Computing How Many Observations Go to Each Value

If k observations take the value α and the remaining $n - k$ observations take the value β, then:

$$k \cdot \alpha + (n - k) \cdot \beta = \text{total}. \tag{3}$$

Solving for k:

$$k = \frac{\text{total} - n \cdot b}{\alpha - \beta}. \tag{4}$$

The function rounds k to the nearest integer and ensures it lies within the range $[1, n - 1]$.

This produces vectors of n generated observations, of which $n - k$ have the value α and k have the value β. This vector of observations represents the distribution of data under these constraints that has the maximum or minimum possible Standard Deviation, respectively.

4. Ensuring Validity on the Discrete Scale

The function verifies that the generated values x_i approximately reproduce the required mean, allowing for rounding:

$$\bar{x}' = \frac{1}{n} \sum_{i=1}^{n} x_i \approx \bar{x}. \tag{6}$$

5. Calculating the Standard Deviation

The sample Standard Deviation is then calculated from the generated observations using the usual formula:

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}. \tag{7}$$

This is done for both the of the generated vectors of data, representing the maximum and minimum Standard Deviation under the constraints. The function returns these min and max SD values to the user.